

A MACHINE LEARNING TECHNIQUE TO IDENTIFY TRANSIT SHAPED SIGNALS

SUSAN E. THOMPSON^{1,2,5}, FERGAL MULLALLY^{2,3}, JEFF COUGHLIN², JESSIE L. CHRISTIANSEN⁴, CHRISTOPHER E. HENZE³,
 MICHAEL R. HAAS³, AND CHRISTOPHER J. BURKE^{2,3}

¹ NASA Ames Research Center, Moffett Field, CA 94035, USA; susan.e.thompson@nasa.gov

² SETI Institute, 189 Bernardo Avenue Suite 100, Mountain View, CA 94043, USA

³ NASA Ames Research Center, Moffett Field, CA 94035, USA

⁴ NASA Exoplanet Science Institute, California Institute of Technology, M/S 100-22, 770 S. Wilson Avenue, Pasadena, CA 91106, USA

Received 2015 July 1; accepted 2015 August 28; published 2015 October 7

ABSTRACT

We describe a new metric that uses machine learning to determine if a periodic signal found in a photometric time series appears to be shaped like the signature of a transiting exoplanet. This metric uses dimensionality reduction and k -nearest neighbors to determine whether a given signal is sufficiently similar to known transits in the same data set. This metric is being used by the *Kepler* Robovetter to determine which signals should be part of the Q1–Q17 DR24 catalog of planetary candidates. The *Kepler* Mission reports roughly 20,000 potential transiting signals with each run of its pipeline, yet only a few thousand appear to be sufficiently transit shaped to be part of the catalog. The other signals tend to be variable stars and instrumental noise. With this metric, we are able to remove more than 90% of the non-transiting signals while retaining more than 99% of the known planet candidates. When tested with injected transits, less than 1% are lost. This metric will enable the *Kepler* mission and future missions looking for transiting planets to rapidly and consistently find the best planetary candidates for follow-up and cataloging.

Key words: binaries: eclipsing – methods: data analysis – planetary systems

Supporting material: machine-readable table

1. INTRODUCTION

As the size and complexity of astronomical data increases, the analysis of these data sets will need to become increasingly automated. When searching for transiting exoplanets, the first step is to test whether the light curve contains a train of transits. Many search algorithms exist to find these periodic events in a time-series of data (Defaÿ et al. 2001; Kovács et al. 2002; Jenkins et al. 2010b; Berta et al. 2012), but the task of selecting which of those truly look transit-like is commonly performed by eye. Exoplanet surveys, such as *CoRoT*, *Kepler*, *TESS*, and *PLATO* (Auvergne et al. 2009; Koch et al. 2010; Rauer et al. 2014; Ricker et al. 2014), have found, or will find, more signals in their data than can easily be examined consistently by a small team of people. While one solution is to enlist citizen scientists such as the planet hunters (Wang et al. 2013), another solution is to develop more sophisticated metrics to cull-out those detections that do not look like transits.

Here we discuss and implement a machine-learning technique to determine whether a signal looks like a transit. This problem can be formulated as a dimensionality reduction and clustering problem. The light curves contain thousands of points that are used to describe the shape of the feature. However, only a few dimensions are needed to describe whether the signal has the steep ingress and egress of a transit, as well as a flat profile outside of the transit event. The trick is to characterize the light curves such that a dimensionality reduction routine clusters signals that look like transits in a region of parameter space separate from those that do not look like transits.

Similar work was done by Matijević et al. (2012) to characterize the eclipsing binaries in the *Kepler* time-series data. Eclipsing binary stars range from detached, with discrete

transit-like dips, to over-contact binaries, that continually vary. Matijević et al. (2012) used an algorithm known as Local Linear Embedding (LLE; Roweis & Saul 2000) to reduce the number of dimensions of the folded light curves of the reported binary stars down to one-dimension. The class of binary star was now mapped onto a continuum of values with detached binaries on one end and over-contact binaries on the other.

The problem we solve in this paper is somewhat different because we attempt to separate transit-like events, i.e., periodic v- or u- shaped variations in the light curve, from all other periodic events found by the *Kepler* pipeline (Jenkins et al. 2010a, 2010b; Wu et al. 2010). The *Kepler* search for transiting planets (Jenkins et al. 2010b) returns transit-like signals as well as other periodic variations. The most common type of false alarm is sinusoidal variations likely caused by spots, pulsations, tidal binaries, or contact binaries. However, the search also returns erratic signals that are likely due to instrumental effects or events that contain no obvious signal at all.

While these signals were removed by hand in previous *Kepler* planet candidate catalogs (Borucki et al. 2011a, 2011b; Batalha et al. 2013; Burke et al. 2014; Mullally et al. 2015; Rowe et al. 2015), the Data Release 24 (DR24) *Kepler* Objects of Interest (KOI) catalog will determine whether a signal is a planetary candidate using an entirely algorithmic approach using a set of algorithms called the Robovetter (J. L. Coughlin et al. 2015, in preparation). The reason for this is driven by the desire to measure accurate planetary occurrence rates (Batalha 2014; Burke et al. 2015; Dressing & Charbonneau 2015). In order to measure the sensitivity of the pipeline to finding planets, the entire search must be done repeatedly on both real and injected signals, demanding automation.

The basic philosophy of the *Kepler* Robovetter is that it uses various metrics to decide if a signal in the data (1) is not transit-like (in shape or significance), (2) has a significant secondary

⁵ a.k.a. Susan E. Mullally.

event (an indication that it is an eclipsing binary), or (3) shows evidence of being due to a background eclipsing binary. To improve the ability of the Robovetter to evaluate the first item, we create a metric that tests whether the shape of the transit is similar to known transiting events.

We describe here a metric that uses the Locality Preserving Projections (LPP; He & Niyogi 2004) dimensionality reduction and k -nearest neighbors to accomplish this task. While we describe the method as it specifically applies to the *Kepler* data, the same technique could easily be adapted to run on the result of any high duty-cycle transit search. In Section 2, we describe the *Kepler* data and the signals found by the pipeline. In Section 3, we describe how we characterize our data, determine a training set, and calculate the LPP transit metric for the Q1–Q17 DR24 KOI catalog. We further evaluate the performance by injecting transits in Section 3.6. Finally, in Section 4, we discuss the performance of the technique on the *Kepler* data.

2. THE *Kepler* DATA

The *Kepler* spacecraft has collected 17 quarters of time series data. Each quarter is approximately 90 days and the cadence of the observations is approximately 29.4 minutes. As a result, *Kepler* has obtained as many as 70,000 brightness measurements of over 160,000 stars listed in the *Kepler* Input Catalog (KIC; Brown et al. 2011) spanning four years (Thompson et al. 2014). The *Kepler* pipeline reduces the data, creates light curves, and searches these light curves for periodic signals that may be consistent with a transit. These signals are known as Threshold Crossing Events (TCEs) and are available at the NASA Exoplanet Science Institute (NExSci) archive⁶ (Akeson et al. 2013). As part of DR24, 20,367 TCEs were discovered by the *Kepler* pipeline (Seader et al. 2015).

The Transit Planet Search (TPS) component of the *Kepler* pipeline performs the search for the transit signals. It does this by whitening the data and searching for significant detections at a large range of periods and using 14 different transit durations. See Seader et al. (2015) and references therein for more information on how TPS searches for transit signals. Once a signal is found, it is sent to the Data Validation (DV) module of the *Kepler* pipeline where it is fit with a transit model and the in-transit points are removed. The same search algorithm is run on the gapped light curve until no more signals are found. In this way, up to 10 TCEs, can be found at different ephemerides on the same *Kepler* target. All of the TCEs and the metrics calculated by these two pipeline modules are available at NExSci.

The technique we describe here only relies on a few of the values calculated by TPS and DV. We use the period, epoch, and duration of the TCE as reported by DV, which are established by fitting the transit model of Mandel & Agol (2002) to the signal. To find the TCE, TPS calculates a Multiple-event Statistic (MES), which gives a measure of the significance of the detected TCE⁷ (Jenkins et al. 2010b). TPS only searches down to an MES value of 7.1;⁸ typically these TCEs appear only marginally above the noise. Sometimes DV fails to fit the signal; in these cases, we revert to the original period and epoch found by TPS and use the pulse duration that

was being used when the signal was originally found. The MES is available regardless of whether the transit model converged.

From these TCEs, the *Kepler* project creates a catalog of planet candidates and astrophysical false alarms (e.g., binary stars and background binary stars), known as the KOI catalog. For every catalog published to date (Borucki et al. 2011a; Batalha et al. 2013; Burke et al. 2014; Mullally et al. 2015; Rowe et al. 2015), all TCEs were examined by a team of astronomers to determine if the data associated with each TCE could potentially be due to a transiting planet (a process known as vetting). Starting with the Q1–Q17 DR24 KOI catalog, this activity is being done by what is being called the “Robovetter” (J. L. Coughlin et al. 2015, in preparation; F. Mullally et al. 2015, in preparation; F. Mullally et al. 2016, in preparation); all tests to determine whether a TCE is a planet candidate will be made by a computer algorithm. This is not an entirely new concept; in previous catalogs some of the tests performed on the transit-like TCEs, specifically to determine whether a transit event occurs on the star in question, were performed by the “Centroid Robovetter” (F. Mullally et al. 2015, in preparation), also McCauliff et al. (2015) implemented a machine learning approach to evaluate the *Kepler* TCEs. For the DR24 KOI catalog, that same philosophy of using metrics and logic is being applied to evaluating the transit shape and whether the transit has a significant secondary eclipse (an indication that it is an eclipsing binary star). The metric we have implemented here focuses on the first question, “Does this TCE look transit-like?.”

3. CALCULATING THE LPP TRANSIT METRIC

In this section, we describe the procedure used to calculate the metric for the DR24 KOI catalog. To summarize, we start with detrended light curves for each TCE. We then fold and bin each light curve into N points. These N points act as the initial number of dimensions that describe each TCE. Using a high signal-to-noise subset of these binned TCEs, we create a map from the initial N dimensions down to smaller n dimensions using the LPP dimensionality reduction algorithm (He & Niyogi 2004). That map is applied to all TCEs. Then, to find the area of this reduced dimensionality space where the transits lie, we create a labeled data set of known transit-like TCEs. For each TCE, we measure the average Euclidean distance to the k nearest transit-like TCEs. This average is the value of the LPP transit metric for the TCE. Each of these steps are discussed in more detail below and the final values for the DR24 TCEs are given in Table 1.

3.1. Detrending the Light Curves

The *Kepler* vetting activity uses two different detrending algorithms to evaluate each TCE. Primarily, it uses the harmonic-removed, median detrender calculated by the DV portion of the *Kepler* pipeline (Wu et al. 2010). In this case, a harmonic series of the largest sine-waves are fit and removed, and then a median detrender is applied with the timescale selected by considering the duration of the signal. We refer to this detrender as the DV-median detrender. The alternate detrender is a non-parametric penalized least-squared (LS) method from Garcia (2010), which includes only the out-of-transit points when computing the filter. Because the in-transit points are not used when detrending, this detrender has the effect of making most signals look more like transits. We refer

⁶ <http://exoplanetarchive.ipac.caltech.edu/>

⁷ MES is a measure of how correlated the data are to a sequence of evenly spaced transit pulses, normalized by the strength of the noise.

⁸ An MES limit of 7.1 ensures only one false alarm due to white noise during the duration of the *Kepler* mission (Jenkins et al. 2002).

Table 1
LPP Transit Metric for DR24 TCEs

TCE (KIC-num)	Period (days)	Duration (hr)	MES	T_{LPP} DV-median	T_{LPP} Penalized-LS	Label
000757450-01	8.88492	2.08	524.0	0.000237	0.000041	TL-C
000892667-01	2.26211	7.51	8.0	0.004608	0.001884	UNK
000892772-01	5.09260	3.40	15.6	0.001337	0.001081	TL
001026032-01	8.46044	4.80	3888.7	0.000466	0.000083	UNK
001026032-02	4.23022	4.61	1439.8	0.000303	0.000039	UNK
001292087-02	1.09524	2.12	9.3	0.002109	0.004084	NOT
001432214-01	161.78830	5.30	839.4	0.000235	0.000038	TL-SS
001432214-02	161.77788	7.60	11.4	0.001103	0.001194	UNK
001434660-02	0.52850	1.04	15.3	0.009662	0.005679	NOT

Note. The TCE period and duration are given for reference, see the NExSci TCE table for full precision of these values. The labels given to the data for training and testing can be interpreted with the following key, see Section 3.4: TL—transit-like, C—candidate, SS—significant secondary, NOT—not-transit-like, UNK—unknown.

(This table is available in its entirety in machine-readable form.)

to this detrender as the penalized-LS detrender. In previous versions of the KOI catalog, members of the team would consider the folded and binned light curve produced by both the DV-median detrender and the penalized-LS detrender to determine if the signal looked sufficiently like a transit. More information about these detrenders and how they have been used to vet planet candidates can be found in Mullally et al. (2015) and Rowe et al. (2015). After detrending, the median value of the light curve is set to zero and the variations are given as fractional changes in the observed brightness.

3.2. Feature Extraction

In order for our metric to examine the shape of each possible transit found in the data, we start by folding the time series and centering the event at a phase of 0.5. We then need to extract, or encode, the shape of this light curve into the same number of data points for all TCEs. To use the dimensionality reduction technique we describe below, all TCEs must start with the same number of dimensions, i.e., data points. Fitting would be an option, as done in Matijević et al. (2012). However, since we are trying to separate non-transit from transiting phenomena, it is not clear what sort of function would easily account for all the different types of variability found among the TCEs as well as a transit shape.

Instead, we bin the folded light curve. The number of bins that optimally describes a light curve depends on the timescale of the interesting phenomena. Most transits look like a rapid dip in the brightness of the star; the transit takes place for a relatively short amount of time compared to the full orbit. For instance, a typical duration for an object in a 20-day orbit is around 10 hr, covering only 2% of the light curve. An Earth transiting a Sun takes ≈ 12 hr; a mere 0.14% of the light curve would be in transit. However, short-period, over-contact eclipsing binaries, stellar spots, and some short-period transiting planets can be fit with a duration lasting upward of 20% of the folded light curve. Since it is the shape of the detected dip in the light curve that we are interested in, and not the duration, we choose our bins in such a way that all transits span approximately the same number of binned points. This is done by using the period and transit duration when picking the range of phases to bin. Also, for the relatively short duration transits, this method has the effect of increasing the importance of the in-transit phases when deciding if the event is transit-like.

However, those signals that confuse transit detectors will similarly fool this metric unless we also consider the light curve at phases away from the transit. For instance, variable stars can look like a broad transit except when you also consider the variability at phases outside of the purported transit event. So, we also include bins at phases away from the transits, and include enough to be able to detect large variability at low harmonics of the detected period.

Finally, in order to encourage our algorithm to discriminate based on the shape of the light curve, and not the amplitude of the signal, we normalize all of our depths to negative one, based on the lowest binned value during the transit.

To summarize, in order to prepare the input matrix to the LPP dimensionality algorithm, we did the following. We started with detrended data, we fold on the TCE period, and then choose approximately one-third of our binned points to cover the phases that lie five transit durations on either side of the reported event. The other two-thirds of our binned points are evenly spaced across phases that do not include those near 0, 0.5, and 1. For the exact way the bins were chosen for the *Kepler* TCEs, see Section 3.5. By not binning points near a phase of 0, we remove the effect of significant secondaries on a number of TCEs with little impact on the measurement of the out-of-transit light curve shape. Consequentially, those TCEs with significant secondaries will still look like transits. The *Kepler* Robovetter has other metrics in place to remove TCEs with significant secondaries. We sort all of the binned values by phase and normalize all of the binned points such that the smallest measured binned value in the in-transit bins has a value of negative one.

We show some examples of binned light curves with transits, using the DV-median detrender, in Figure 1 and non-transiting binned light curves in Figure 2. Notice that for long-duration, short-period TCEs, the two sets of binned points overlap in phase. For both figures the input to the dimensionality algorithm is shown on the right; the phases are sorted and the depth is normalized. As a result of this binning procedure, the general shape of the transit is preserved but information about the depth and duration of the transit have been removed.

To give an idea of how much the detrenders can disagree on the shape of a TCE, we show in Figure 3 the binned points for the same TCEs with the different detrenders. In the first four panels, we specifically pick cases where the detrenders disagree so much that it significantly changes the outcome of the LPP

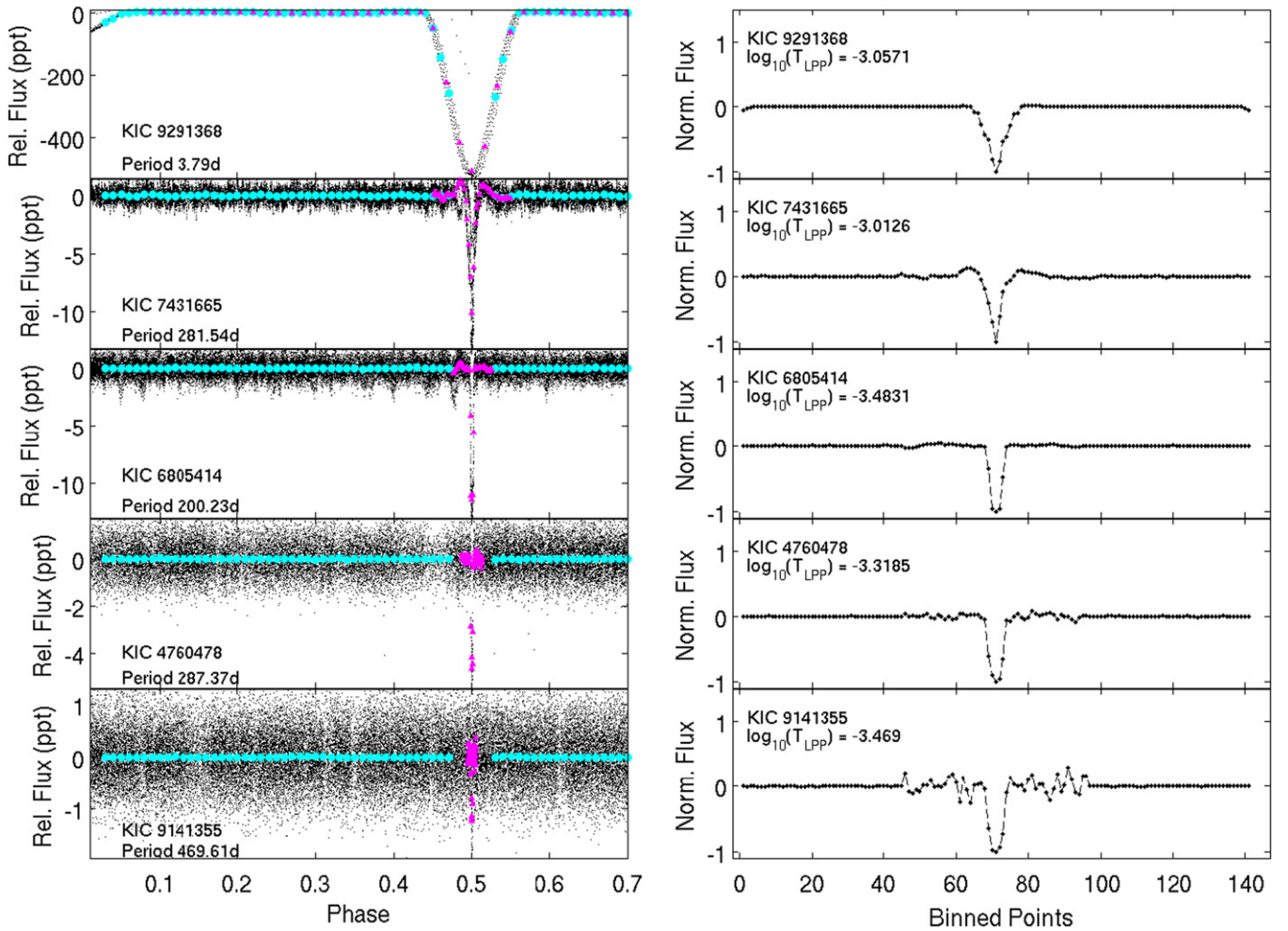


Figure 1. Left: the folded light curves for five example TCEs (black) along with the binning in units of parts per thousand. Those binned points that span five times the transit duration are shown with magenta triangles and those that cover all but the in-transit points are shown with cyan circles. Right: the normalized binned points sent to the LPP algorithm. This method of preparing the data has the effect of making all transits have the same depth and similar widths. Because the width of the in-transit bins can be much smaller, the binned points on the right can appear much noisier near the transit.

transit metric’s value. In these cases, the light curve is extremely variable; how much of that variability is removed by the detrender will drastically change the appearance of the light curve, especially when it is folded and binned.

3.3. Dimensionality Reduction

Because we are trying to reduce the complex information of a binned light curve down to the simple question of whether it looks like a transit or not, we considered the power of dimensionality reduction. While many of these routines, which attempt to maintain the most diverse dimensions in a data set could work for our purposes, we settled on LPP⁹ (He & Niyogi 2004) as implemented by the MATLAB Toolbox of Dimensionality Reduction¹⁰ (van der Maaten et al. 2009). One advantage of LPP over nonlinear algorithms, such as LLE, is that the mapping is unambiguously defined everywhere in the higher dimensionality space; and so LPP can be precisely evaluated on injected signals, even though it was trained on only the original sample.

LPP is similar to the most commonly used linear dimensionality reduction algorithm, Principle Component Analysis (PCA). While PCA attempts to maximize the variance of the data in each dimension, LPP attempts to preserve the local neighborhoods, as measured by a Euclidean k -nearest-neighbors, when reducing the dimensions. In this way, LPP has the advantage of being less sensitive to outliers than PCA and also has the ability to map highly nonlinear manifolds as can be done by nonlinear techniques. For more information on precisely how this algorithm works, see the [appendix](#) and He & Niyogi (2004). While the linearity and locality aspects of LPP are useful in the problem we are trying to solve, ultimately, we chose this particular dimensionality reduction algorithm because it did a good job of preserving the transit-shape information in binned light curves.

The LPP algorithm takes our N binned points for a sample of the TCEs, and maps them to a lower number of dimensions, n , attempting to preserve those elements of the set that are adjacent to each other, as defined by using k -nearest neighbors. Once our TCEs are represented by the lower number of dimensions, we calculate the distance to an integer number, k , of nearest known transits from the labeled data set described below. We use MATLAB’s *knnsearch* algorithm using a

⁹ <http://papers.nips.cc/paper/2359-locality-preserving-projections.pdf>

¹⁰ <http://lvdmaaten.github.io/drtoolbox>

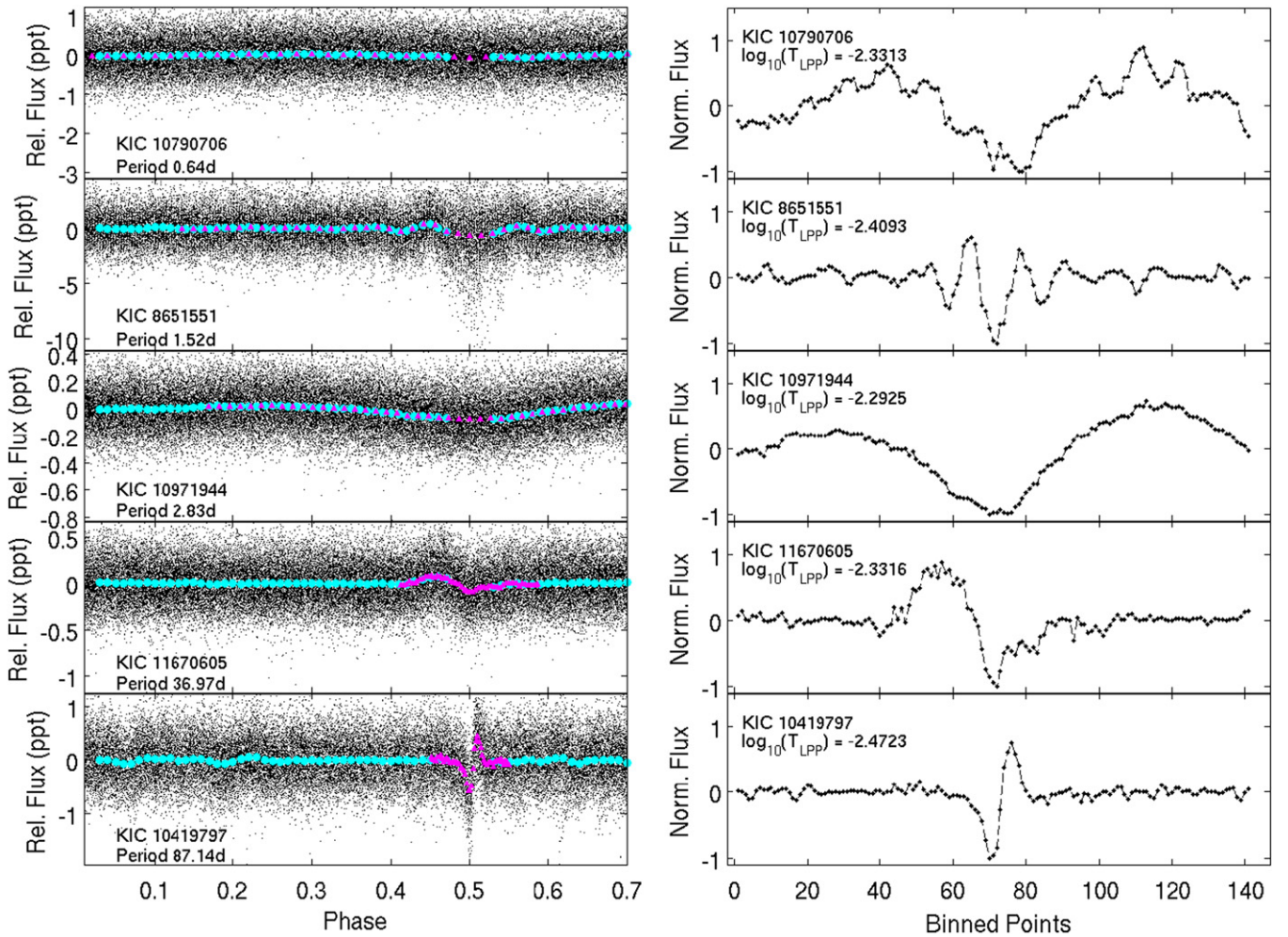


Figure 2. Left: the folded light curves of TCEs that are likely not transiting planets (black) in units of parts per thousand. Both groups of binning are shown, those that span five times the transit duration are shown with magenta triangles, and those that cover all but the in-transit points are shown with cyan circles. Right: the normalized binned points sent to the LPP algorithm.

Euclidean distance (in MATLAB, this is performed using a Minkowski distance with an exponent set to two). The mean of these k distances is what we report as the LPP transit metric. For ease, we use the symbol T_{LPP} to represent the LPP transit metric. If the TCE is as close to the known transits as the known transits are to each other, the TCE is deemed to be transit-like.

3.4. Creating a Labeled Data Set

The key to the success of this metric is having a good set of known transit-like signals for training. In this case, we create a labeled data set from those TCEs previously found and vetted by the *Kepler* Project (Batalha et al. 2013; Burke et al. 2014; Mullally et al. 2015; Rowe et al. 2015). Those TCEs whose ephemeris matched those of known KOIs in the cumulative table at the NExScI archive (except for those marked with the “not-transit-like” flag) became the sample of known transiting objects. Note that KOIs that have dispositions of false positive due to being a centroid offset or because of an ephemeris match are also included as transit-like objects in our labeled data set.

For testing purposes, we also federated the TCE sample with those TCEs from earlier catalogs (specifically Q1–Q16 and Q1–Q12) that previously failed to become KOIs. These objects were all evaluated by individuals and deemed either to look

not-transit-like or were too low signal-to-noise to be made into a KOI. Those non-KOI objects that matched the ephemeris of our Q1–Q17 TCE list were labeled as not-transit-like.

We separately track two groups within the transit-like population. First, we track those objects that are known to be planetary candidates in the cumulative KOI table at NExScI. These make a very high fidelity population of objects known to have a transit shape. Second, we track those objects with a known secondary eclipse, and likely to be eclipsing binaries by using the false positive flag available at NExScI. The intent is to keep the eclipsing binaries at this stage in the vetting because many of them look sufficiently like transiting planets. Other metrics are used by the *Kepler* Robovetter to remove these from the planet candidate sample. Thus, those objects with known significant secondaries are counted as transit-like.

In total, we label 5678 TCEs as transit-like, of which 3738 are planetary candidates, 633 have a significant secondary, and 1307 are other known false positives, and we label 1039 as not-transit-like. The entire transit-like population is important because that is what defines the parameter space that should contain transit-like TCEs. Note that we set aside 10% of the transit-like set for testing purposes and did not use them to train our metric. Table 1 provides the labels given to each TCE for training and testing purposes.

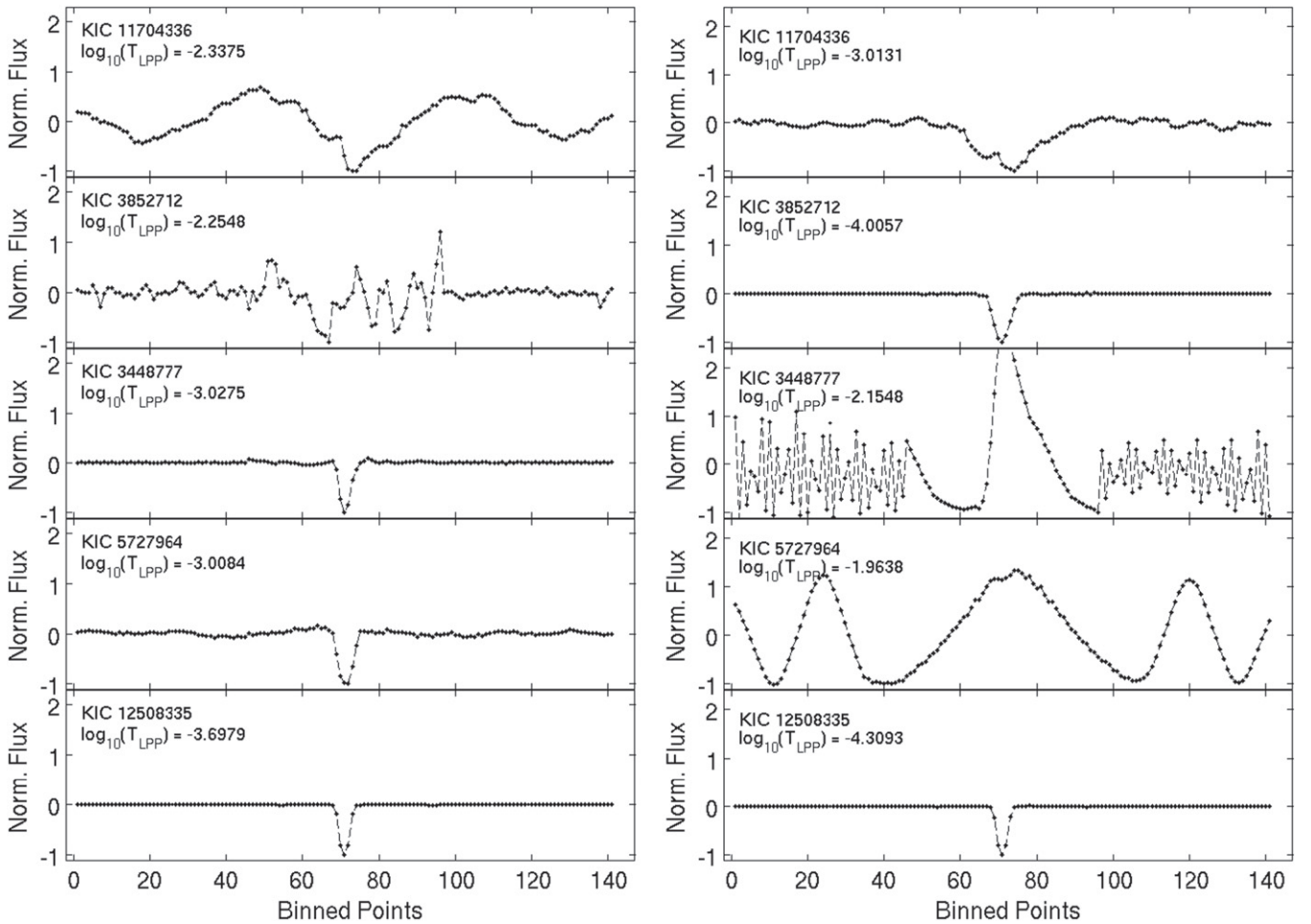


Figure 3. Binned points for the same TCEs as seen by the two different detrenders. The DV-median detrender is on the left and the penalized-LS detrender is on the right. The first four panels show TCEs where the detrenders disagree and the last case is an example of where they agree. The $\log_{10}(T_{LPP})$ value is given on the plot below the KIC number. Typically, a value larger than -2.5 indicates a non-transit-like shape.

There are a few small issues with creating the labeled data set in this way. Some of the low signal-to-noise objects from previous TCE catalogs will now be legitimate transit-like signals because more transits are available and because of improvements to the pipeline that remove known noise sources. As a result, we might not expect all of the not-transit-like TCEs to fail our metric. Also, the false positive flags available at the archive were not universally set for every vetted KOI. The largest impact of this is that a very small population of not-transit-like KOIs are masquerading as transit-like objects (likely less than 1% of the transit-like objects). Also, some known binaries may not be included in the significant secondary set. These are issues we keep in mind when using the labeled data set for training and testing our metric.

3.5. Evaluating the Q1–Q17 DR24 TCEs

We apply the above technique to calculate T_{LPP} for every Q1–Q17 DR24 TCE and give the values in Table 1. While there are several tuneable parameters in creating the LPP transit metric (e.g., k , N , n , phase span of the in transit bins, etc.), we discovered that changing these parameters does not drastically change the outcome. The values chosen for these parameters were determined empirically by trying to remove the most known non-transit like signals while keeping the most transit like signals.

For this implementation, which is being used by the DR24 *Kepler* Robovetter (J. L. Coughlin et al. 2015, in preparation), we started with $N = 141$ binned data points, 51 in-transit and 90 out-of-transit. This was chosen so that the spacing of the out-of-transit bins is 0.01 in phase and the spacing of the in-transit bins is 0.2 times the duration in phase. The out-of-transit bins were evenly spaced across the phases 0.03–0.47 and 0.53–0.97. For most TCEs, this results in the bin spacing of the in-transit bins being smaller than the bin spacing of the out-of-transit bins. We then used LPP to reduce the dimensions of the data set to $n = 20$, where the TCE locality was determined by a k -nearest neighbor test with $k = 15$. We used all TCEs with an $MES > 8$ to create the mapping to the lower dimension, thereby removing those signals that were likely dominated by noise. For every TCE, we reduce the dimensionality with this mapping and measure the distance to the $k = 15$ closest known transits. We provide the T_{LPP} values for both detrenders in Table 1.

While 20 dimensions are hard to display simultaneously, we can show the final metric and how it performed on the labeled TCEs. Figure 4 shows the histogram of $\log_{10}(T_{LPP})$ calculated for the labeled data set. There is significant separation in those objects that are known planet candidates from those that appear not-transit-like for the DV-median detrended data. One way to naively pick a value that could act as a cutoff line between

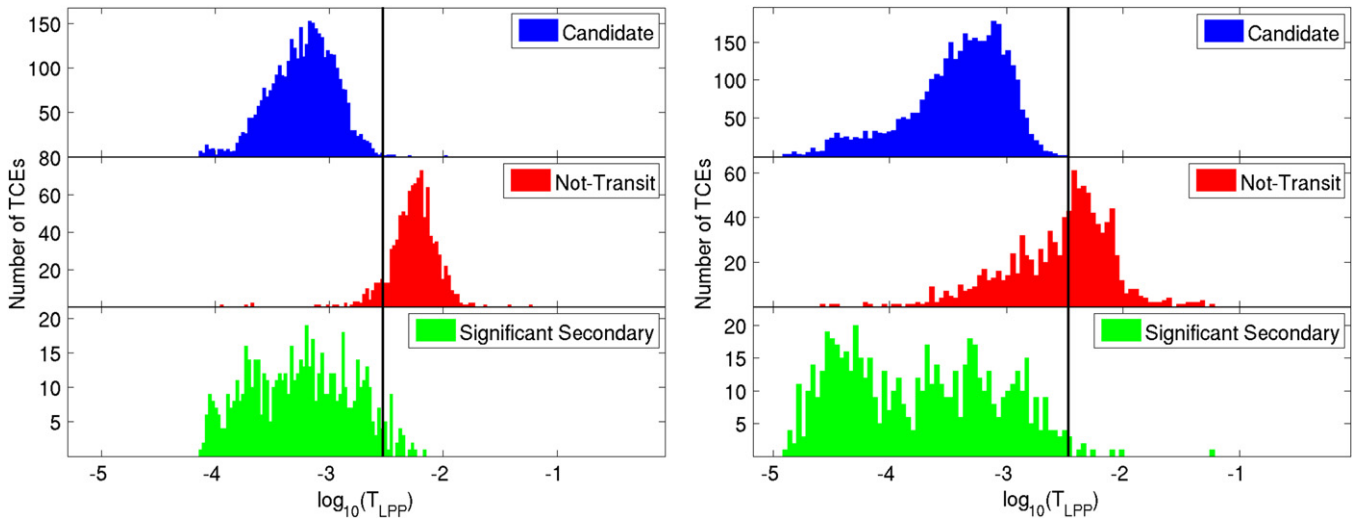


Figure 4. Histogram of the LPP transit metric for the labeled data set using the DV median (left) and penalized-LS (right) detrenders. A three sigma line discussed in the text is shown in black to help guide the eye. Note that the transit-like population has been split into candidates and significant secondaries to show the slightly higher failure rate among binaries.

transit-like and not-transit-like would be three times the standard deviation of T_{LPP} for the known transit-like distribution. For the DV-median detrender this line would be $T_{LPP} = 0.003001$ ($\log_{10}(T_{LPP}) = -2.52$). If this line is chosen, T_{LPP} rejects 0.3% of the candidates, 4.4% of those with significant secondaries, and 91.2% of the not-transit-like objects. For the test set, the metric fails $<0.1\%$ and 6.7% of the candidate and binary-like sets, respectively. For the unlabeled data set, using this cutoff rejects 75.1% of the TCEs.

We show the same histograms for the light curves detrended by the penalized-LS in Figure 4. Notice the much poorer separation between the candidates and not-transit-like data sets. This detrender preserves the location of the known transit when detrending and as a result many variable star signals look like transits. If we draw the same three sigma line based on the distribution of known transit-like events ($T_{LPP} = 0.00344$, $\log_{10}(T_{LPP}) = -2.463$), we reject only 51.7% of the known not-transit-like TCEs while preserving $>99.9\%$ of the transit-like and 98.6% of the binary TCEs. We get almost identical values for the test set. As a result, the penalized-LS T_{LPP} value should only be used to supplement the results of the DV-median T_{LPP} value, especially if the goal is to remove the most false signals from the catalog.

3.5.1. Code Availability

The MATLAB code, as it was run to produce the LPP transit metric for the Q1–Q17 DR24 TCE list and the transit injection run, is available on SourceForge at <http://sourceforge.net/projects/lpptransitlikemetric>. It includes the code to bin the light curves and create the transit metric from the training set.

3.6. Transit Injection

As a second test of the LPP transit metric, we apply it to transits that are injected into the light curves. This pixel-level transit injection is performed in order to understand what transit signals are not found by the *Kepler* pipeline, a necessary step to calculate the occurrence rate of planets (Christiansen et al. 2013). The transits were injected into all 17 quarters of data, but not with the expected distribution of the detected planets.

Instead, they were injected with longer periods and at lower MES to carefully probe the parameter space where the *Kepler* pipeline is less likely to find planets. See Christiansen et al. (2013) for more information on transit injection for the *Kepler* pipeline. We used a 17 quarter transit injection run to test our metric. It produced 11,326 injected TCEs and used the same procedures as that discussed in Christiansen (2015) and Christiansen et al. (2013), but this is not the final transit injection run being used to calculate the average detection probability or test the Robovetter (Christiansen et al. 2015; J. L. Coughlin et al. 2015, in preparation). This preliminary run was sufficient for our purposes of showing how well the metric preserves known transiting phenomena.

We use the procedure above to calculate T_{LPP} for the injected transits using both detrenders. As stated before, we can easily apply our previous map (created from the original real TCEs) to the injected TCEs. Similarly, when we apply the k -nearest neighbors, we are measuring the distance to the original set of known transit-like TCEs, not to the injected TCEs. In Figure 5, we show a histogram of the LPP transit metric for the 11,326 injected transits for both sets of detrended light curves. Notice that, in general, the lower MES transits have larger values for T_{LPP} , meaning they are less well-separated from the non-transiting phenomena. This indicates that users of this metric may want to use a pass/fail threshold value that is dependent on MES in order to eliminate a larger fraction of the higher MES not-transit-like events without removing more borderline, low-MES events.

Using the same value established above to indicate the line between transit-like and not-transit-like, we reject 94, or 0.8%, of the injected transits for the DV-detrender and 23, or 0.2%, of the injected transits for the penalized-LS detrending. This is a slightly higher, though entirely acceptable, rate of failure for this metric when compared to the training set statistics given in Section 3.5.

When inspecting those injected transits incorrectly classified by the DV-median metric as not-transit-like, we see that they are predominantly of low MES; 2% of injected TCEs with recovered MES < 10 fail the metric. Since low MES objects are predominantly transits of small radii, this introduces a very

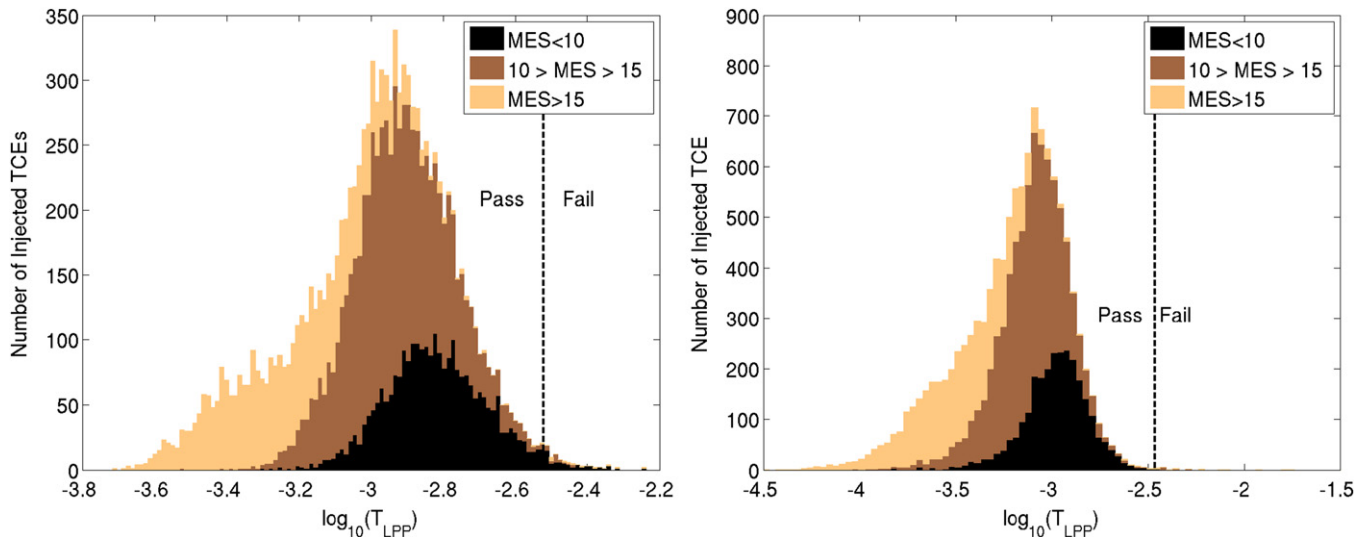


Figure 5. Histogram of the LPP transit metric for the injected transits using the DV-median (left) and penalized-LS (right) detrenders. The histogram was calculated and stacked for different ranges of measured MES. Using the cutoff established by the real TCEs of $\log(T_{\text{LPP}}) = -2.5$, only a fraction of a percent of injected TCEs would be lost.

slight bias toward incorrectly classifying the signals of some of the smallest planets. This is not true for the penalized-LS detrender, where the minimal tail is distributed more evenly across MES. Some of the higher MES injections that fail the metric are caused by injections onto highly variable stars and the detrenders did not preserve the shape of the transit. Also, there is no trend between the fraction of rejected signals as a function of period, even as the threshold is relaxed. A larger investigation of the detection efficiency across various parameters using the entire Robovetter is available in the DR24 KOI catalog paper (J. L. Coughlin et al. 2015, in preparation) and with the detection efficiency products available at NExSci (Coughlin 2015).

4. DISCUSSION

The LPP transits metric is being used by the DR24 *Kepler* Robovetter (J. L. Coughlin et al. 2015, in preparation) to decide whether a TCE is a planetary candidate. The Robovetter is responsible for choosing a pass/fail threshold. In the past, these catalogs have erred on the side of keeping questionable transits at the expense of retaining more false alarms. Even when this metric is implemented with a single, conservative value, as demonstrated above, it eliminates a large fraction of not-transit-like TCEs. Primarily, it does a good job of removing the short-period, long-duration TCEs that look like sine-waves, not transits. The final results from the *Kepler* Robovetter and how it uses all of its metrics, including this one, will be available at NExSci.

Another automated transit finding technique is known as the Autovetter (Jenkins et al. 2014; McCauliff et al. 2015). The Autovetter uses a random forest, machine-learning technique to decide which TCEs are planetary candidates. For the Q1–Q17 DR24 TCE autovetter catalog (Catanzarite et al. 2015), the inputs now include the DV-median LPP transit metric. In this case, the random forest uses a training set (which is created in a manner very similar to how we created our training set) to learn how important T_{LPP} is when deciding on whether a TCE is a transiting event. The full results of the Autovetter’s vetting will be found in the TCE table at NExSci for Q1–Q17 DR24. For

this run, the LPP transit metric is ranked as the second most important metric for making its decisions (Catanzarite et al. 2015).

While there are several ways that we can change the value of T_{LPP} for each TCE, by far the most important is how the data is prepared before applying the LPP dimensionality reduction. The number of chosen nearest neighbors and the number of reduced dimensions are relatively inconsequential in comparison to how the data is binned and how the data is detrended. This is exemplified in the performance for two different detrenders; the DV-median detrender does a far better job of preserving the differences between the transit-like and not-transit-like light curves. However, there are times when the penalized-LS detrending is more accurate.

For example, one known issue with using this metric occurs for transits found on variable stars, especially if that variability is near to a harmonic of the period of the transit. These can be classified as not-transit-like because the detrending significantly distorts the signal. Note that this effect could similarly fool the astronomers who did the manual vetting in previous catalogs. This is the reason two different detrendings were made available for the vetting activity. In those cases where the harmonic removal step of the DV-median detrender removed or distorted the transit, usually the penalized-LS would preserve it. One way to improve our results would be to add some intelligence to the *Kepler* Robovetter to help it pick the more accurate detrenders.

There are several populations of binary systems that are known to have higher failure rates with this metric. Eclipsing binaries with large eccentricities and a large secondary eclipse get flagged as not-transit-like because of the large deviation from a flat continuum at out-of-transit phases. Also, eclipsing binaries make-up a smaller portion of the transit-like training set, and thus, it is more difficult for the deeper binaries to have 15 nearby neighbors. This problem could be mitigated by using the *Kepler* Eclipsing Binary Catalog to add binaries to the transit-like training set (Prša et al. 2011; Slawson et al. 2011). Finally, the Heartbeat stars, a class of dynamically distorted binary systems (Thompson et al. 2012), are classified as transit-like at times because the signal can include a large, discrete

negative deviation in the brightness. However, Heartbeat stars are rare (<160 are known in the *Kepler* field) and most known Heartbeat stars do not create TCEs ($<10\%$ in Q1–Q17 DR24).

5. CONCLUSIONS

We present a new metric to discriminate between signals found by the *Kepler* Pipeline that look like transits and those that do not. By folding and binning the light curve and applying machine learning techniques that rely on our knowledge about what a transit signal looks like in the *Kepler* data, we calculate the LPP transit metric. This metric is able to remove over 90% of the known not-transit-like TCEs while preserving over 99% of the known planet candidates and over 99% of injected transit signals. As currently implemented, this metric will prevent hundreds of not-transit-like events from populating the Q1–Q17 DR24 KOI table at the expense of losing only a few real transiting events. The Robovetter has additional metrics that will weed out some of the remaining signals, improving the reliability of the final catalog.

This LPP transit metric makes its decisions on TCEs consistently, reliably, and rapidly, a large improvement over the manual activity performed by teams of astronomers in the past. Because *Kepler* plans to calculate planetary occurrence rates, it is necessary to have an estimate of the planet catalog's completeness and reliability. Using metrics like these, the Robovetter can quickly create a new catalog after any improvement to the pipeline, and can also run on injected signals. The latter makes it possible to more easily calculate the sensitivity of the pipeline across planet type for occurrence rate calculations.

Such a metric can easily be extended to work with current missions such as *K2* (Howell et al. 2014; Foreman-Mackey et al. 2015), and future exoplanet missions such as the *Transiting Exoplanet Survey Satellite* mission (*TESS*, Ricker et al. 2014) and the *Planetary Transits and Oscillations of Stars* mission (*PLATO*, Rauer et al. 2014) to quickly identify the best transiting planets. Metrics that can quickly and consistently evaluate potential signals in the data are even more important for these future missions because the data volume is even larger than *Kepler* and rapid vetting of the detected signals will enable immediate follow-up observations of the best planet candidates.

We thank the larger *Kepler* team for their support and hard work in making this data available and in supporting this paper. Funding for the *Kepler* mission is provided by the NASA Science Mission Directorate. We also thank the referee for useful and insightful comments that improved the clarity of the manuscript. Some of the data presented in this paper were obtained from the Multi-mission Archive at the Space Telescope Science Institute (MAST). STScI is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS5-26555. Support for MAST for non-*HST* data is provided by the NASA Office of Space Science via grant NNX09AF08G and by other grants and contracts. This research has made use of the NASA Exoplanet Archive, which is operated by the California Institute of Technology, under contract with the National Aeronautics and Space Administration under the Exoplanet Exploration Program.

APPENDIX LPP: A SUMMARY OF THE ALGORITHM

Here we summarize the basic logic of how LPP calculates its transformation matrix. For a more mathematically complete description of the LPP algorithm consult He & Niyogi (2004).

The idea of LPP is to map a data vector x_i (i.e., a folded, binned light curve) with N dimensions to y_i with n reduced dimensions, i.e., $y_i = A^T x_i$, where A^T is an $n \times N$ transformation matrix. A is a matrix of eigenvectors calculated by minimizing the distance between reduced dimension vectors, y , only if they were near each other in the higher dimensions.

LPP considers only the nearest neighbors in its calculations by constructing a symmetric weighting matrix, W_{ij} , that has a value of one when two data points are connected and zero otherwise. Data vectors i and j are considered connected to each other if i is among the k nearest neighbors of j or if j is among the k nearest neighbors of i . In this way, the number of connections for any one vector must be at least k , but can be larger when other vectors also consider i to be a nearest neighbor. Using this weighting matrix, the function that is minimized when choosing the eigenvectors, is as follows:

$$\sum_{ij} (y_i - y_j)^2 W_{ij}. \quad (1)$$

Here, i and j run over the full number of data vectors (i.e., binned TCE light curves) in the data set. Because W_{ij} only contains 1's when two x vectors are adjacent, this function will be large when similar looking binned light curves are mapped far from each other. However, non-neighboring vectors do not factor into the minimization. Thus, when LPP minimizes the above equation to find the transformation matrix, it preserves the local structure of the data. The eigenvectors with the smallest eigenvalues (A) are used as the transformation matrix to map x_i to y_i . This transformation matrix may then be applied to any vector in the higher dimensions to reduce its dimensions.

In comparison, PCA maximizes the variance in the mapped data using the function:

$$\sum_i (y_i - \bar{y})^2. \quad (2)$$

Because LPP only considers local distances, outliers will not weigh as heavily into the calculation and thus LPP can be more robust to outliers than PCA. Also, because it tries to keep similar data vectors together, it can have better discriminating power than PCA, especially when the data vectors cluster into distinct groups in the higher dimensionality space.

REFERENCES

- Akeson, R. L., Chen, X., Ciardi, D., et al. 2013, *PASP*, **125**, 989
- Auvergne, M., Bodin, P., Boissard, L., et al. 2009, *A&A*, **506**, 411
- Batalha, N. M. 2014, *PNAS*, **111**, 12647
- Batalha, N. M., Rowe, J. F., Bryson, S. T., et al. 2013, *ApJS*, **204**, 24
- Berta, Z. K., Irwin, J., Chabonneau, D., et al. 2012, *AJ*, **144**, 145
- Borucki, W. J., Koch, D. G., Basri, G., et al. 2011a, *ApJ*, **728**, 117
- Borucki, W. J., Koch, D. G., Basri, G., et al. 2011b, *ApJ*, **736**, 19
- Brown, T. M., Latham, D. W., Everett, M. E., & Esquerdo, G. A. 2011, *AJ*, **142**, 112
- Burke, C. J., Bryson, S. T., Mullally, F., et al. 2014, *ApJS*, **210**, 19
- Burke, C. J., Christiansen, J. L., Mullally, F., et al. 2015, *ApJ*, in press, arXiv:1506.04175
- Coughlin, J. L. 2015, KSCI-19096-001
- Catanzarite, J., et al. 2015, KSCI-19091-001
- Christiansen, J. L. 2015, KSCI-19094-001

- Christiansen, J. L., Clarke, B. D., Burke, C. J., et al. 2013, *ApJS*, **207**, 35
- Christiansen, J. L., Clarke, B. D., Burke, C. J., et al. 2015, *ApJ*, in press
- Defay, C., Deleuil, M., & Barge, P. 2001, *A&A*, **365**, 330
- Dressing, C. D., & Charbonneau, D. 2015, arXiv:1501.01623
- Foreman-Mackey, D., Montet, B. T., Hogg, D. W., et al. 2015, arXiv:1502.04715
- Garcia, D. 2010, *Computational Statistics and Data Analysis*, **54**, 1167
- He, X., & Niyogi, P. 2004, *Advances in Neural Information Processing Systems*, **16**, 37
- Howell, S. B., Sobeck, C., Haas, M., et al. 2014, *PASP*, **126**, 398
- Jenkins, J. M., Caldwell, D. A., Borucki, W. J., et al. 2002, *ApJ*, **564**, 495
- Jenkins, J. M., Caldwell, D. A., Chandrasekaran, H., et al. 2010a, *ApJL*, **713**, L87
- Jenkins, J. M., Chandrasekaran, H., McCauliff, S. D., et al. 2010b, *Proc. SPIE*, **7740**, 0
- Jenkins, J. M., McCauliff, S., Burke, C., et al. 2014, in *IAU Symp. 293, Auto-Vetting Transiting Planet Candidates Identified by the Kepler Pipeline*, ed. N. Haghighipour, 94
- Koch, D. G., Borucki, W. J., Basri, G., et al. 2010, *ApJL*, **713**, L79
- Kovács, G., Zucker, S., & Mazeh, T. 2002, *A&A*, **391**, 369
- Mandel, K., & Agol, E. 2002, *ApJL*, **580**, L171
- Matijević, G., Prša, A., Orosz, J. A., et al. 2012, *AJ*, **143**, 123
- McCauliff, S. D., Jenkins, J. M., Catanzarite, J., et al. 2015, *ApJ*, **806**, 6
- Mullally, F., Coughlin, J. L., Thompson, S. E., et al. 2015, *ApJS*, **217**, 31
- Prša, A., Batalha, N., Slawson, R. W., et al. 2011, *AJ*, **141**, 83
- Rauer, H., Catala, C., Aerts, C., et al. 2014, *ExA*, **38**, 249
- Ricker, G. R., Winn, J. N., Vanderspek, R., et al. 2014, *Proc. SPIE*, **9143**, 20
- Rowe, J. F., Coughlin, J. L., Antoci, V., et al. 2015, *ApJS*, **217**, 16
- Roweis, S. T., & Saul, L. K. 2000, *Sci*, **290**, 2323
- Seader, S., Jenkins, J. M., Tenenbaum, P., et al. 2015, *ApJS*, **217**, 18
- Slawson, R. W., Prša, A., Welsh, W. F., et al. 2011, *AJ*, **142**, 160
- Thompson, S. E., Caldwell, D., Jenkins, J., et al. 2014, KSCI-19064-002
- Thompson, S. E., Everett, M., Mullally, F., et al. 2012, *ApJ*, **753**, 86
- van der Maaten, L. J. P., Postma, E. O., & van den Herik, H. J. 2009, *Tilburg University Technical Report, TiCC-TR 2009-005*
- Wang, J., Fischer, D. A., Barclay, T., et al. 2013, *ApJ*, **776**, 10
- Wu, H., Twicken, J. D., Tenenbaum, P., et al. 2010, *Proc. SPIE*, **7740**, 19